In *Proceedings of NLPRS'97, Natural Language Processing Pacific Rim Symposium*, pages 95-100, Phuket, Thailand.

Topic Graph Generation for Query Navigation: Use of Frequency Classes for Topic Extraction

Yoshiki Niwa, Shingo Nishioka, Makoto Iwayama, and **Akihiko Takano**
Advanced Research Laboratory, Hitachi, Ltd.
Hatoyama, Saitama 350-03 Japan
{*yniwa, nis, iwayama, takano*}*@harl.hitachi.co.jp*

Yoshihiko Nitta
Dept. of Economics, Nihon University
1-3-2 Misaki-chô, Chiyoda-ku, Tokyo 101 Japan
nitta@eco.nihon-u.ac.jp

Abstract

To make an interactive guidance mechanism for document retrieval systems, we developed a user-interface which presents users the visualized map of topics at each stage of retrieval process. Topic words are automatically extracted by frequency analysis and the strength of the relationships between topic words is measured by their co-occurrence. A major factor affecting a user's impression of a given topic word graph is the balance between common topic words and specific topic words. By using frequency classes for topic word extraction, we made it possible to select well-balanced set of topic words, and to adjust the balance of common and specific topic words.

1 Introduction

As the rapid development of networks brings us easy access to large quantities of on-line information, the need to find ways of retrieving useful information is increasing. Retrieval is not always an easy task, however, because an inquiry that is imagined by a person cannot be directly conveyed to a machine, and because inquiries themselves are not always clear. A guidance function which supports an interactive approach to the target information is therefore required, and wwe have developed a guidance system that presents to users a visualized map of topics at each stage of interaction.

This is done by automatically constructing graphs of topic words in the retrieved documents. Topic words are extracted by frequency analysis and the strength of the relationships between topic words is measured by their co-occurrence.

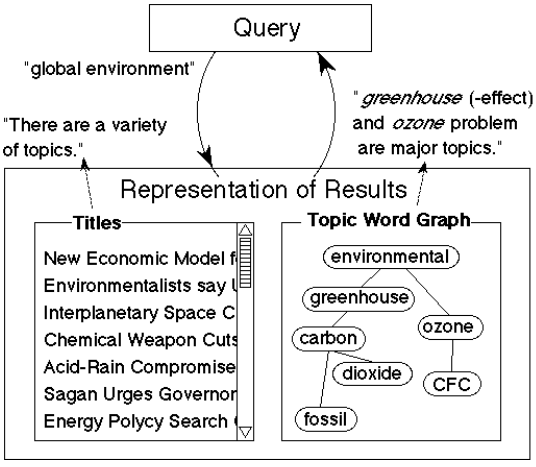


Fig. 1. Interactive IR

Figure 1 illustrates the flow of our interactive retrieval system. After a user starts with a query, a set of documents is retrieved and displayed as a list of their titles. From these retrieved documents are extracted topic words along with their mutual relationships, and these words and relationships are displayed in a graph structure. From the title list we can get concrete information about the retrieved results; for example, "documents concering such and such topics were retrieved." From the topic graph, on the other hand, we get more abstract information like "such and such topics are dominant in the retrieved documents." By referring to both concrete and abstract information

concerning the results, users can proceed to the next trial with a better perspective.

2 Related Work

Visualization of information has been attracting a lot of interest, and many studies have been done (Rao et al. (1995)). Some of them visualize the document space and others visualize the lexical space. Our study does the latter. A closely related work is the scatter-gather method developed by Cutting *et al.* (1992). In their method, retrieved documents are automatically clustered according to the similarity of their document vectors, then the characteristic words of each cluster are extracted. One problem with this method is the computational complexity of clustering, which is about the third power of the number of documents. In the development of our system, we paid attention to the real-time response and therefore to the computational tractability.

Another active issue in lexical visualization is the merger of a thesaurus database and automatically extracted topic words. This issue has been studied by the Illinois University group in their digital library project (Johnson and Cochrane, 1995; Schatz, Johnson, and Cochrane, 1996). The creation of additional effects beyond the sum of two factors may be a future problem.

Visualization of document spaces seems to be more popular than that of lexical space. In the Smart system (Salton et al., 1994), text relation maps are constructed according to the similarity of their vector representation. Xerox's Information Visualizer system (Mackinlay, Rao, and Card, 1995), visualizes the reference structure of documents. One major problem with the visualization of document space is the representation of documents. To give proper information to users, document titles need to be presented. However, they are not always compact. This is where the visualization of lexical space is advantageous.

In Japan, visualization of lexical space has primarily been studied in the framework of idea processing systems by the group at Tokyo University (Sumi et al., 1995). Recently, Sugimoto (1996) proposed a method of displaying documents and related words at the same time. Interaction of documents and related words should be a goal for the future research. There are also some research studies concerning the document browsing system using the visualization technique (Morohashi et al., 1995; Arita, Yasui, and ichiro Tsudaka, 1995).

3 Generation of Topic-Word Graphs

Our topic graph generation method consists of following three steps.

1. Topic word extraction.
2. Link generation by co-occurrence analysis for extracted topic words.
3. Graph mapping to a 2-dimensional area.

An outline of the method has already been described (Niwa, Iwayama, and Takano, 1997). Here we explain the details with reference to Fig. 2.

3.1 Topic word extraction

Topic word extraction is based on the importance of each word appearing in the retrieved set of documents. We measure the importance of each word by the ratio of df to DF, where df is the number of retrieved documents containing the word and DF is the number of all documents containing the word in the entire target database. We call this ratio the relative frequency of the word. This importance measurement is an embodiment of a natural idea that if a word not so common in general appears frequently in the retrieved results, then the word is probably characteristic to the retrieved set of documents.

Figure 2 shows the case where the query is *global environment*. The word *greenhouse*, for example, appears in 62 documents in the retrieval results and in 268 documents in the whole database. So the relative frequency is $62/268 = 0.23$. We calculate the importance (relative frequency) of all words appearing at least once in the retrieved documents and select some of the most important words.

Relative frequency is an intuitively acceptable measurement of importance and has the additional merit of not being affected very much by random sampling of retrieved documents. It has a problem, however, in that it suffers from noises caused by low frequency words. The improvement of this simple method is given in section 4.

3.2 Link generation by co-occurrence analysis

In our method, each topic word X is linked to a word which has the highest co-occurrence strength

with respect to X among the topic words having higher frequency than X. The co-occurrence strength of Y with respect to X is measured by the number of documents containing both X and Y by the number of documents containing Y. In case of *ozone*, the words having higher frequency are *global*, *greenhouse*, and *dioxide*. Of these three words, *dioxide* has the highest co-occurrence strength with respect to *ozone*, 0.40. Therefore, *ozone* is linked to *dioxide*. By applying this process to all topic words, we get the link table (at the right of the co-occurrence table).

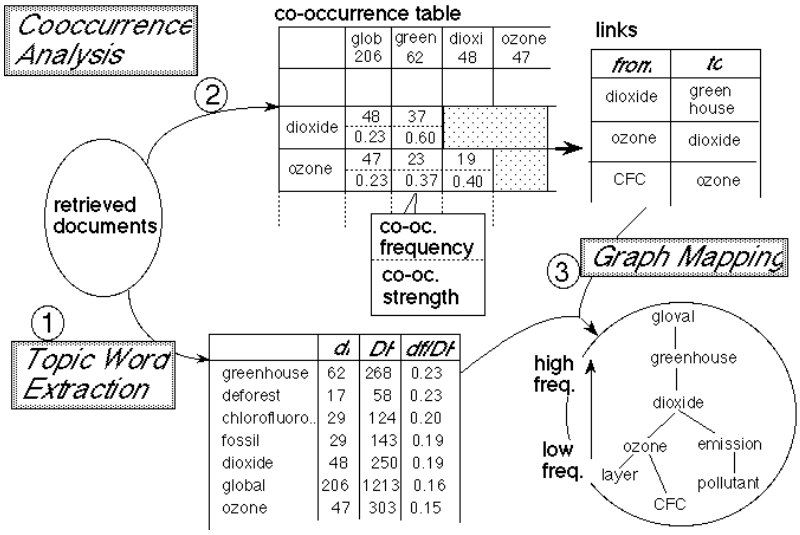


Fig. 2. TWG Generation Method

3.3 Graph mapping

In the first step we got a set of topic words which form the nodes of a topic word graph to be generated, and in the second step we got links between these topic words. Thus we now have the topological structure (nodes and links) of the topic word graph. In this step we map the graph to a 2-dimensional rectangular area of given size.

The mapping is done by determining the xy-coordinates of each node. Again, there are many possible methods. In what follows we describe the method currently used in our prototype system.

The y-coordinate of a node is first derived from the document frequency (df) of the corresponding topic word by using the following formula: $y = C_1 \ \tan^{-1}(C_2 \ log\,(df\,/\,df_m))$., where, df_m is the middle frequency of all of the topic words.

The terms C_1 and C_2 are constants. Then x-coordinates are allocated recursively starting from the top frequency node, just to prevent overlaps.

4 Topic-word extraction using frequency classes

The relative frequency of a word is a natural index for measuring how prominent a word is in a particular set of documents. If topic words are selected according only to this relative frequency, however, we usually miss some important high-frequency words.

When the topic words of a retrieved document set are displayed for characterizing the document set, the balance of common and specific topic words is important. This balancing problem cannot be solved by using other criterion for measuring the importance. For example, the *tf*idf* measurement is advantageous for selecting high frequency words but misses some important specific topic words. We therefore used a topic word extraction method based on the classification of words by their document frequency (number of documents containing a word) in the retrieval results.

In the framework of our topic word extraction method using frequency classes, we use the following parameters for setting the frequency thresholds defining the frequency classes and the allotment of number of topic words to each frequency class.

- N Total number of topic words to be extracted
- C Number of frequency classes
- L The lower frequency boundary
- b Balance parameter

4.1 Thresholds of frequency classes

The first stage is the definition of frequency classes, and the parameters C and L are used for setting the frequency thresholds. The lower frequency boundary L takes a value within the interval (0 1], and words less frequent than $L \times M$ are excluded from the candidates of topic words. Here M is the maximum document frequency of a word in the retrieved documents over the set of all words appearing in the retrieved documents.

The frequency classes are determined by parameters C and L in the following way. Here df means the document frequency of a word.

$$\text{Class } k : M\ r^k \leq df < M\ r^{k-1},$$

$$\left(r = max\left(L, \frac{1}{M}\right)^{1/C} \right).$$

(As an exception, $df = M$ is classified into class 1.)

4.2 Balance Tuning

Another advante of using frequency classes is that we can change the balance of high-frequency and low-frequency topic words. The balance parameter b, which takes a value within the interval [-1, 1], is used for this purpose. The upper limit $N_b(k)$ of the number of topic words taken from frequency classes 1 through k is determined by

$$N_b(k) = N \times \left\{ b \times \left(\frac{k}{C}\right)^2 + (1-b) \times \left(\frac{k}{C}\right) \right\}.$$

When this balance parameter takes a negative value (such as -1), that means higher weights are attached to high-frequency words (or the words in the frequency classes of small class numbers), and when the parameter takes a positive value (such as 1), higher weights are attached to the low-frequency words (or the words in the frequency classes of large class numbers).

In the above definition we determined the number of topic words taken from each frequency class indirectly using the accumulated number over classes 1 to k. The reason we did not determine the number directly for individual classes is that in some cases some frequency classes do not contain enough words. In such cases the above definition allows us to make a compensation by taking extra topic words from the next frequency class.

4.3 Example

A primary benefit of using frequency classes in topic word extraction is that we can adjust the balance of common topic words and specific topic words, and this balance affects the impression made by topic word graphs.

Figure 3 shows six different topic word graphs generated from a single set of documents retrieved by using *ASEAN*. as the query. The six cases are as follows.

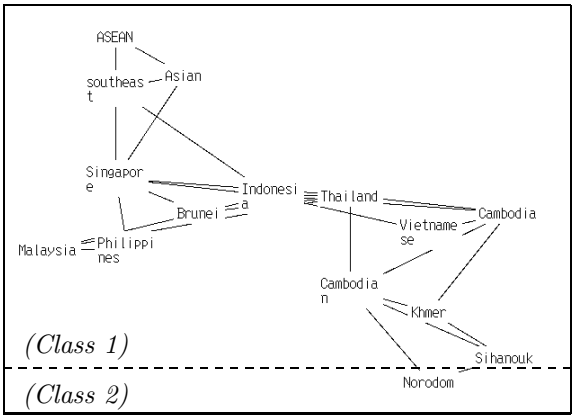
(a) No use of frequency classes.
(b) Only frequency class 1 is used.
(c) Frequency classes 1 and 2 are used.
(d) Frequency classes 1 through 3 are used.
(d') Same as (d), with balance parameter b= -1.0.
(d") Same as (d), with balance parameter b= +1.0.

Let us first examine the topic word graphs for cases (a) and (b), respectively the case where frequency classes are not used and the case where only frequency class 1 is used. These two graphs are very similar, with the small difference that *Nordom* in case (a) is replaced by *Rouge* in case (b). For comparison, the frequency threshold (freq.= 19) between classes 1 and 2 is shown in (a) by a dotted horizontal line. Since the line is almost at the bottom of the graph, we know that if we do not use frequency classes most of the topic words are taken from frequency class 1. In other words, the case of no use of frequency classes can be well simulated by the case in which only high frequency classes are used.

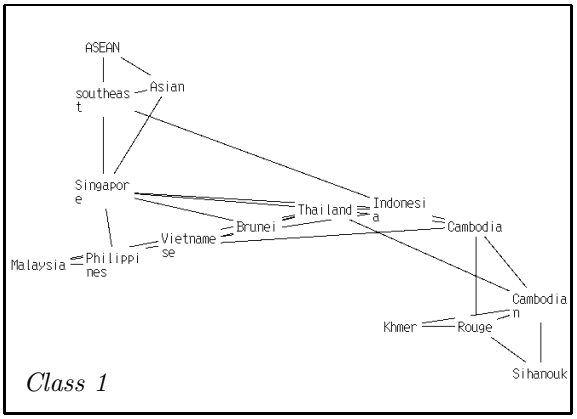
Graph (c) is the topic word graph generated when topic words are taken from the upper two classes, 1 and 2. Since the balance parameter b is set to 0 (neutral), almost same numbers of topic words are taken from both classes. (Seven from class 1 and eight from class 2.) The dotted horizontal line in the middle of the graph shows the frequency threshold (freq.=19). Since the vertical coordinate is proportional to the logarithm of frequency, the threshould line is located almost at the center. The words belonging to class 1 are *ASEAN* (the query word), and some related country names such as *Thailand* and *Singapore*, and this selection is similar to the case in cases (a) and (b). Some words related to Cambodia, however, such as *Cambodia* and *Khmer Rouge*, which belong to class 1 in case (b), are substituted by other Cambodia-related words, such as *Hun Sen* or *Phnom Pehn*, belonging to class 2. This means that topic words related to a specific topic, in this case Cambodia rather than the general topic ASEAN, are more likely to appear in case (b) than in case (a).

The graph (d) is topic word graph when the three frequency classes 1 to 3 are used. Here again, the balance parameter is set to 0 (neutral),

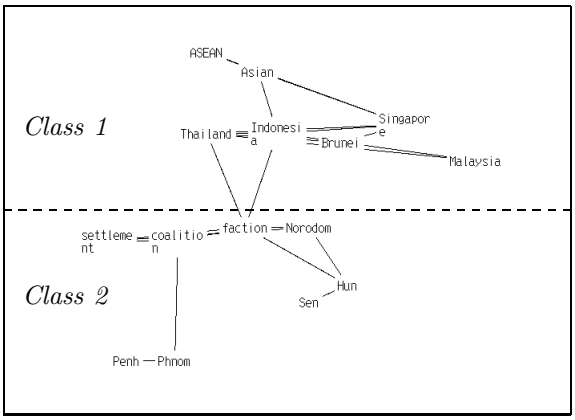
(a) No use of freq. classes



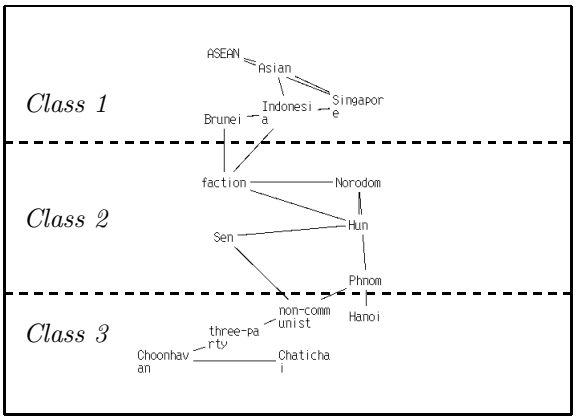
(b) Only freq. class 1 is used



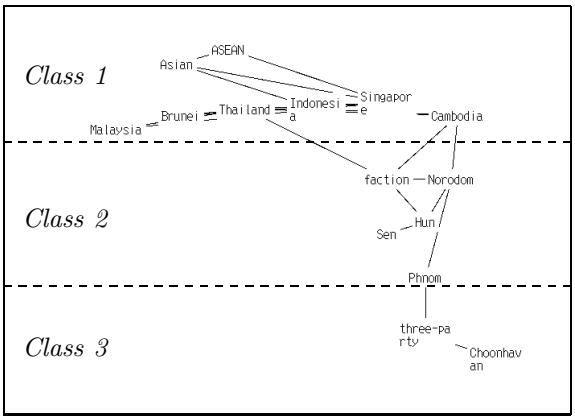
(c) Freq. classes 1 and 2 are used



(d) Freq. classes 1 through 3 are used



(d') Classes $1 \sim 3$, with $b = -1.0$



(d") Classes $1 \sim 3$, $b = +1.0$

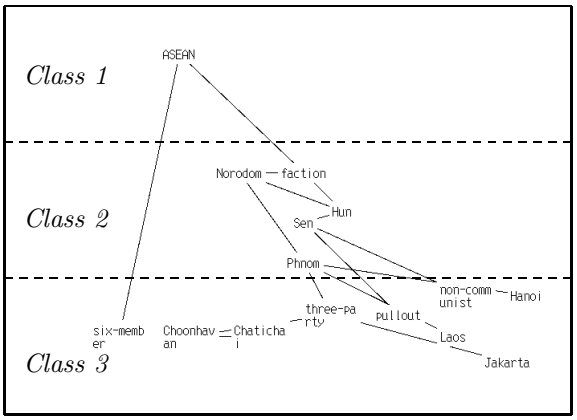


Fig. 3. Effect of the use of frequency classes

and each five words are taken from three classes. This time, five new words in class 3 – *Hanoi*, *non-communist*, *three-party*, and *Chatichai Choonhavan* (the then prime minister of Thailand) – are added in place of *Thailand* and *Malaysia* in class 1 and *settlement*, *coalition*, and *Penh* in class 2. This change is interpreted as showing that we get still-more-specific words related to a major specific topic rather than a general topic.

The cases (d') and (d") are same as case (d) with respect to the frequency classes but with different balance parameters. In case (d') the balance parameter b is set to -1.0 and more weight is on the common words in class 1. Actually, the numbers of topic words taken from classes 1 to 3 are 8, 5, and 2 respectively. Since the weight of class 3 is very small, the graph is similar to the case (c) (or even (a) or (b)). Conversely, in case (d") the balance parameter is set to 1.0 and more weight is on the specific words in class 3. This time, nine words are taken from class 3, 5 words from class 2 and only 1 word from class 1. Since most of words are taken from class 3, the topic word graph seems to outline a major specific problem rather than whole articles concerning *ASEAN*. (In this case the major specific problem is the peace talks at Jakarta by Cambodia's Hun Sen government and three guerrilla groups.)

5 Conclusion

To make an interactive guidance mechanism for document retrieval systems, we developed a user-interface which presents users a visualized graph of topic words at each stage of the retrieval process. Topic words are automatically extracted by frequency analysis and the relationship between topic words is measured by their co-occurrence.

We built a prototype retrieval system for about 80 thousand articles of AP Newswire '89, and experiments with this system support our expectation that the guidance provided by the topic word graph is useful for interactive screening. By using the topic word extraction method using frequency classes, we succeeded in taking well-balanced topic words from the wide range of word frequencies. This method is also advantageous for adjusting the balance of the high-frequency topic words and low-frequency topic words, a balance that greatly affects the user's impression of the topic word graph.

References

Arita, Hidekazu, Terumasa Yasui, and Shin ichiro Tsudaka. 1995. Information strolling through automatically organized information space. *IPSJ SIG Notes*, 95-NL-108:69–74. (In Japanese).

Cutting, Douglass R., David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Ann Int'l SIGIR'92*, pages 318–329.

Johnson, Eric H. and Pauline A. Cochrane. 1995. A hypertextual interface for a searcher's thesaurus. In *Proceedings of the Digital Libraries '95*, Austin, Texas. Center for the Study of Digital Libraries, Texas A&M University.

Mackinlay, Jock D., Ramana Rao, and Stuart K. Card. 1995. An organic user interface for searching citation links. In *Proceedings of ACM CHI'95*, pages 67–73, Denver, Colorado. ACM.

Morohashi, Masayuki, Koichi Takeda, Hiroshi Nomiyama, and Hiroshi Maruyama. 1995. Information outlining - filling the gap between visualization and navigation in digital libraries. In *Proceedings of International Symposium on Digital Libraries*, pages 151–158, Tsukuba.

Niwa, Yoshiki, Makoto Iwayama, and Akihiko Takano. 1997. Interactive support of query refinement by dynamic word co-occurrence. In *Proceedings of ICCPOL'97*, pages 383–386.

Rao, Ramana, Jan O. Pedersen, Marti A. Hearst, Jock D. Mackinlay, Stuart K. Card, Larry Masinter, Per-Kristian Halvorsen, and George G. Robertson. 1995. Rich interaction in the digital library. *Communications of the ACM*, 38(4):29–39.

Salton, Gerard, James Allan, Chris Buckley, and Amit Singhal. 1994. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264:1421–1426, June.

Schatz, Bruce R., Eric H. Johnson, and Pauline A. Cochrane. 1996. Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. In *Proceedings of ACM DL'96*, page (to appear), Bethesda, Maryland. ACM.

Sugimoto, Masanori, Teruo Koyama, Koichi Hori, Setsuo Ohsuga, Hiroshi Kinukawa, and Hisao Mase. 1996. A retrieval system for visualizing the relations between documents. *IPSJ SIG Notes*, 96-NL-112:15–22. (In Japanese).

Sumi, Yasuyuki, Ryuta Ogawa, Koichi Hori, Setsuo Ohsuga, and Kenji Mase. 1995. Css: A human communication support system by visualizing thought space structure. *Technical Report of IEICE*, TL95(6):11–22. (In Japanese).